Networks as representations of complex data

Gesine Reinert

Department of Statistics
University of Oxford
reinert@stats.ox.ac.uk

Workshop on the Analysis of Complex Multi-Level Data
Oxford, November 17, 2015

What are networks?

Networks are just graphs; they are described by a set of nodes (vertices) and a set of edges (links) between nodes.

They are often described using summary statistics such as the average degree. The degree of a node is the number of other nodes to which it is connected by an edge.

The density of a network is the number of edges divided by the possible number of edges.

The global clustering coefficient of a network is the number of triangles / (number of 2-stars + triangles)

Here are some examples of networks (graphs).

Marriage relations between Florentine families

Networks as representations of complex data

Gesine Reinert

What are networks?

Typical network observations

Where are we now?

Summary

Yeast protein-protein interaction networks

Networks as representations of complex data

Gesine Reinert

What are networks?

Typical network observations

Where are we now?

Summary

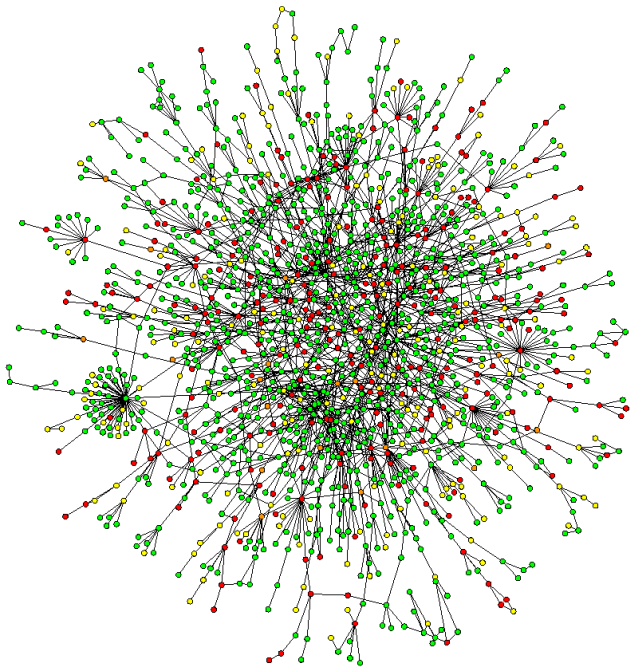


Figure : From *Yeong et al. (2001)*. Proteins are nodes, and two nodes are connected if a physical interactions between the protein has been observed. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown).

KEGG pathway for Parkinson's disease

Networks as representations of complex data

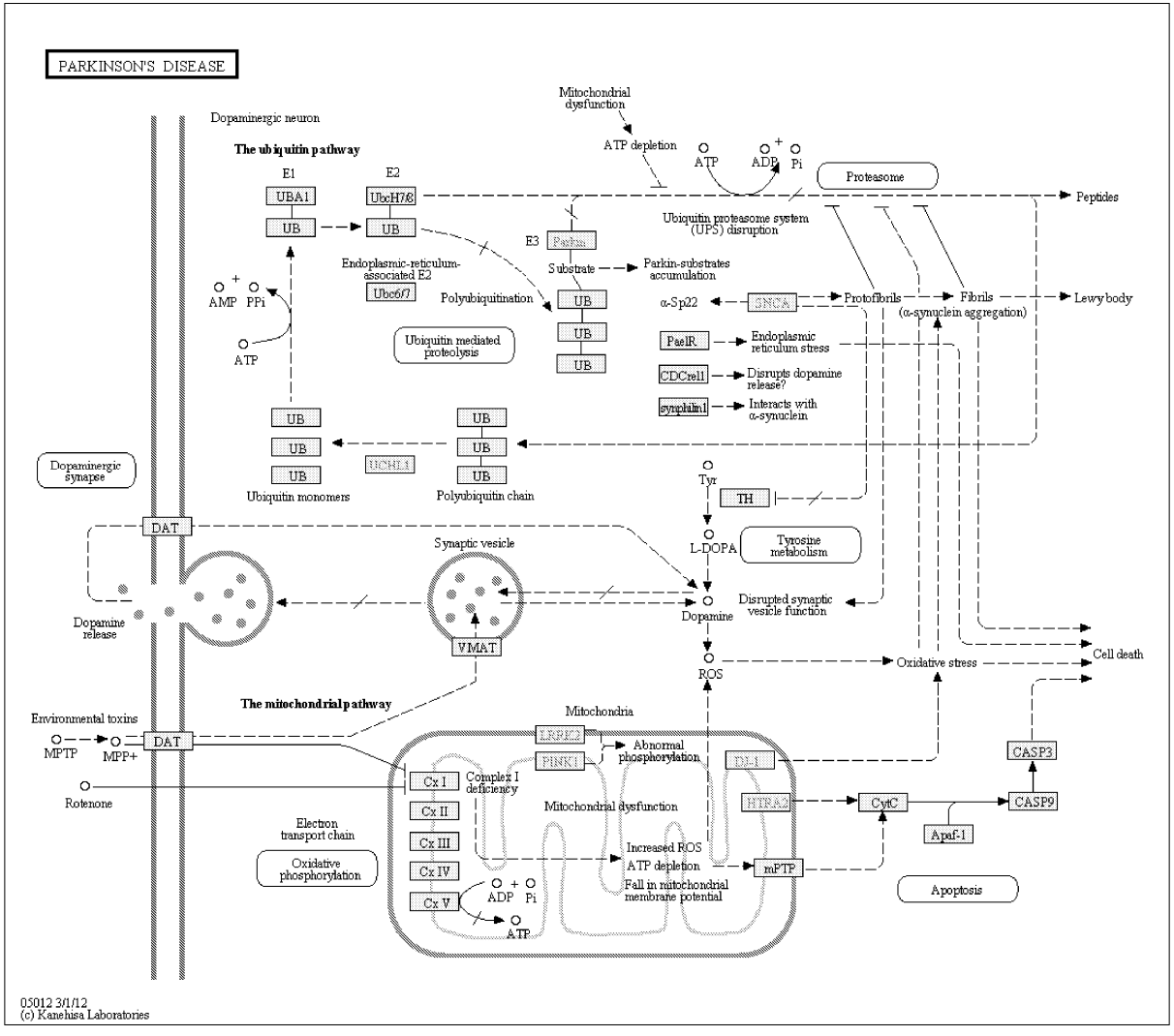
Gesine Reinert

What are networks?

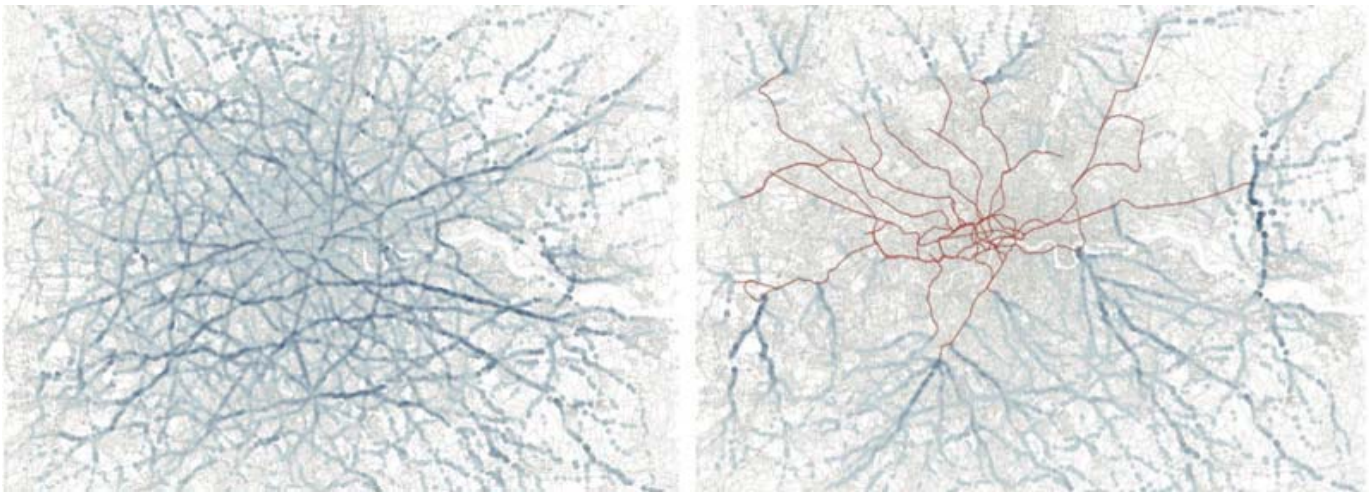
Typical network observations

Where are we now?

Summary



London congestion (*Spano et al. 2015*)



Betweenness centrality.
Left panel: underground system operates at the same speed as the average speed on roads.
Right panel: underground system operates at 10 times the speed as the average speed on roads. The city centre is relieved; congestion occurs at the terminal stations of underground lines.

Networks arise in a multitude of contexts.

Networks for environmental factors could be

- friendship/family networks, including communities such as schools
- geographical networks,
- epidemic spread networks,
- travel networks,
- power grid,
- mobile phone networks,
- sampling design (main street bias)
- ...

Typical networks in systems biology are

- *Protein-protein interaction (PPI) networks*: no dynamics, not spatial; high false-positive and false-negative rates.
- *Gene regulatory networks*: The activity of genes is regulated by transcription factors: proteins that typically bind to DNA. Bipartite network of transcription factors and binding sites.
- *Metabolic networks*: The chemical compounds of a living cell are connected by biochemical reactions which convert one compound into another. Reaction network of compounds; edges are reactions (typically).
- *Cell signaling networks*: Cell signaling pathways interact with one another to form networks; these typically integrate PPI networks, gene regulatory networks, and metabolic networks.

Networks are increasingly used as representations of complex data sets such as gene expression data and social interaction data. Edges often reflect interdependence relations.

The number of nodes varies from less than 100 (social networks) to 4.73 billion pages (the Internet, Thursday, 12 November, 2015) and beyond.

From a theoretical viewpoint networks are classified as dense, sparse, or moderately sparse.

For dense networks from exchangeable models there is a limit theory available (Aldous, Hoover, Lovasz, Janson, Bollobas, Diaconis)

Many real-world networks are sparse (c.f. Dunbar number).

What do we observe in networks?

In real networks, given the number of nodes and the number of edges, compared to the Bernoulli random graph model where all edges are equally likely and occur independently of each other, we usually see

- a higher clustering coefficient than expected;
- a shorter average shortest path length than expected;
- more nodes with high degrees than expected (heavy tails).

The phenomenon of short paths, often coupled with high clustering coefficient, is called the *small world phenomenon*.

A typical situation

Find relationships between networks which are

- too large for eyeballing;
- with a large number of mis-specified edges;
- for which no good parametric model is available with regards to the research question.

Big questions

- Informative network summaries, even when the network is not fully known;
- Network models which capture dependence between edges, and inference for such models;
- Interplay between network architecture and efficiency / robustness;
- Identification of relevant substructures (communities, modules) ;
- Comparison of networks which are of different size and which may contain different nodes, yet which are hypothesized to be related;
- Integration of different interlinked networks.

Where are we now? Methodological:

Networks as representations of complex data

Gesine Reinert

What are networks?

Typical network observations

Where are we now?

Summary

- Informative network summaries, even when the network is not fully known: sample small subnetworks to estimate local summaries: *Holmes and R. (2004), Bhattacharyya and Bickel (2013)*
- Network models which capture dependence between edges, and inference for such models: stochastic blockmodel inference: *Airoldi, Costa and Chan (2013), Latouche and Robin (2013), Wolfe and Olhede (2013),* shortest paths asymptotics *Barbour and R. (2012)*
- Interplay between network architecture and efficiency / robustness: percolation on networks, *v.d. Hofstad (2014)*

- Identification of relevant substructures (communities, modules): Algorithms based on *Newman-Girvan (2003)*, overlapping communities
  statistical analysis: *Bickel and Chen (2009)*
- Comparison of networks: Comparison based on subgraphs and subsampling *Przulj (2006), Ali et al. (2014)*
- Integration of different interlinked networks: Multilayer networks: *Bianconi (2015); Kivel'a et al. (2014)*

Where are we now? Some examples

- *Huitsing et al. (2014)* Victims, bullies and their defenders: studies the coevolution of positive and negative networks, based on questionnaire data.

  They use social network models to analyse the interplay between defender and bullying networks, with focus on exogeneous variables (age, gender) as well as network structure (reciprocal relations, triadic relations).

- *Handcock et al. (2015)* use respondent-driven sampling to estimate the size of two sub-populations which are most at risk for HIV in two cities in El Salvador (overall HIV prevalence estimated at 0.8%).

  Respondent-driven sampling is a method network sampling, in which subsequent sample members are selected from among the social relations of current sample members. Respondents choose which of their contacts will be sampled next.

  The analysis is Bayesian and incorporates degrees of subjects.

- *Denon et al. (2013)* Social contacts on given days for more than 5,000 respondents in England, Scotland and Wales: online and questionnaire data.

  Children, public-sector and healthcare workers have the highest number of contact hours and are therefore most likely to catch and transmit infectious disease.

  Clustering varies across social settings and increases with duration, frequency of contact and distance from home.

  Age is an important indicator of social mixing patterns and daily contacts. School-age and pre-school children are associated with the greatest contact times, as well as the greated number of contacts.

- *Ghiassian et al. (2015)* A DIseAse MOdule Detection algorithm: this study analyses protein-protein interaction networks from 70 complex diseases: 141,296 interactions betweeen 13,460 proteins, 1,531 of which are associated with one or more diseases.

  Using a version of community detection the authors identify the full disease "modules" around a set of known disease proteins.

  In total, 58 of the 70 modules can be validated either by functional annotation or by pathways, 46 by both.

Summary: how can networks be relevant?

Networks as representations of complex data

Gesine Reinert

What are networks?

Typical network observations

Where are we now?

Summary

Networks are a useful data structure for unordered data of dependent observations such as:

- social networks;
- networks for the spread of disease;
- cellular networks.

Networks are a very active area of research.

Many problems are still to be solved!